Pathogenesis related proteins: A defensin for plants but an allergen for humans

Rahul Arora, Amit Kumar, Indrakant Kumar Singh, Archana Singh

PII: S0141-8130(19)33972-8

DOI: <https://doi.org/10.1016/j.ijbiomac.2019.11.223>

Reference: BIOMAC 14004

To appear in: *International Journal of Biological Macromolecules*

Received date: 28 May 2019

Revised date: 10 September 2019

Accepted date: 27 November 2019

Please cite this article as: R. Arora, A. Kumar, I.K. Singh, et al., Pathogenesis related proteins: A defensin for plants but an allergen for humans, *International Journal of Biological Macromolecules*(2019), <https://doi.org/10.1016/j.ijbiomac.2019.11.223>

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Pathogenesis Related Proteins: a defensin for plants but an allergen for humans

Rahul Arora^{a,b,c}, Amit Kumar^a, Indrakant Kumar Singh^{*d,e} Archana Singh^{*a,e}

^a Department of Botany, Hansraj College, University of Delhi, Delhi 110007, India.

^b Francis Crick Institute, 1 Midland Road, London NW1 1AT, United Kingdom

^c Division of Biosciences, University College London, London WC1E 6BT, United Kingdom

^d Molecular Biology Research Lab, Department of Zoology, Deshbandhu College, University of Delhi, Kalkaji, New Delhi 110019, India

^e Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Beutenberg Campus, Hans-Knöll-Str. 8, D-07745 Jena, Germany.

*Corresponding authors:

E-mail addresses: archanasingh@hrc.du.ac.in (A. Singh), asingh@ice.mpg.de;

iksingh@db.du.ac.in, (I.K. Singh), isingh@ice.mpg.de

Highlights

1. Out of 19 classes of Pathogenesis-related (PR) proteins, eight are recognized potent allergens.
2. Using *in silico* approaches, we tested and confirmed that other eleven families of PR-proteins are also potential allergen.
3. This analysis will aid in identification of allergenicity caused due to PR-proteins and to avoid clinical complications.

Abstract

Pathogenesis related (PR) proteins are defensins expressed upon biotic and abiotic stress by plants. They are also synthesized in specific plant parts such as root, stem, leaves, pollen grains and fruits. According to the current classification, there are 19 different classes of PR-Proteins, 8 out of 19 are proven to exhibit allergic reactions in human. Here, we carried out the allergenicity potential test for the remaining 11 classes of PR-proteins using *in-silico* approaches. Our analysis suggests that the other 11 families also have the allergenic potential. We modelled the proteins for which our predictions suggested them to be allergens and for which the crystal structures were not available. We then predicted the B-cell epitope binding regions for all the proteins and used molecular docking approach to study the allergen-antibody interaction. Our findings suggest that all the 11 protein families analysed, can potentially be classified to be as allergens. We also provide evidence that the number of IgE binding-residues in the allergens is correlated with their respective binding energies with the IgE molecule. This study will be highly relevant to understand the allergenic potential of PR-proteins to make an informed decision about the consumption of food with high degree of PR-expression.

Introduction

Plants synthesize a class of defense-associated proteins upon attack by pathogens, known as 'Pathogenesis-Related' Proteins or the PR-Proteins [1]. They provide a fine line of defense against a great variety of pathogens but are also expressed in plants upon abiotic stress such as cold, drought, salt and heavy metal stress and hence are also known as the stress proteins [2]. These proteins were initially classified into 17 distinct classes but 2 new members (PR-18 and PR-19) have been added to the list. Out of these 19 families, PR-1, PR-2, PR-3, PR-4, PR-5, PR-8, PR-10 and PR-14 have already been classified as allergens. The previously known proteins as allergens have been classified under two categories of allergens: airborne and food allergens depending upon their mode of entry into the human body [3]. These proteins are known to trigger allergic symptoms, such as contact dermatitis, eye and airway allergy, asthma, rhino conjunctivitis and oral allergy syndrome which is characterised by symptoms like: swelling of lips, palate, tongue; feeling of nausea and the sensation of tightness. If these allergens have been consumed in greater amount, the gastrointestinal symptoms are also triggered [4,5,6]. Apart from these proteins triggering specific allergic symptoms associated with them, they also cross-react and can

worsen the symptoms. An example of such a case is seen when PR-3 cross-reacts with PR-4 thus, triggering symptoms associated with the latex-fruit syndrome; similarly the pollen allergens of Mugwort and Birch cross-reacts with food allergens from celery and spices such as cumin, thyme and Oregano leading to a disease known as mugwort-birch-celery-spice syndrome that mimics oral allergy and gastrointestinal symptoms [7,8](Figure 1).

Since, these proteins can trigger a diversity of allergenic responses, therefore the prediction, verification and validation of their allergenic potential is an important domain to investigate. Various prediction methods have been proposed to classify proteins as allergens or non-allergens that use different algorithms, such as machine learning, sequence alignment search method, epitope mapping techniques etc. The specificity and sensitivity of the predictions depends upon the algorithm employed for prediction. Apart from their allergenicity potential prediction, the B-cell epitope binding residue predictions for these allergens are also of significant importance. B-cell epitope binding residue prediction is based on the fact that B-cells express B-cell Receptors (BCRs) on their surface, which consists of the membrane-bound Immunoglobulin (Ig) molecules in complex with other proteins and they also contain a complementarity determining region (CDR) which interacts with the allergen [9,10]. The B-cell epitope prediction tools uses physico-chemical properties of the proteins such as residue accessibility, its hydrophilic nature and the structural properties such as the location of an amino acid in loops, sheets or helices to predict the allergen-antibody interaction [9]. Therefore, the protein should have the correct 3D structure. The structure of these proteins is derived experimentally but if it is not possible to purify the crystals, they are often modelled using the approaches such as homology and remote homology modelling, threading approach, and *ab-initio* method [11].

Molecular docking of these allergens with the antibody can provide a better understanding about the interactions that takes place between these molecules at the atomic levels. Docking is a method that represents the various orientations of the interacting molecules in which they are most stably bound [12]. The results of these stable orientations are often accompanied with an associated energy value which is a sum total of various atomic forces, such as van der Waals forces, electrostatic forces, solvation potentials, steric forces along with the hydrogen bonding and the partial charge distribution on amino acid residues [13]. Therefore, molecular docking is a promising method that can be used to study such protein-protein interactions. In this study, our analysis focuses on

figuring out the status of 11 proteins of the PR-family with unknown allergenicity potential using various allergenicity prediction tools and B-cell epitope predictions followed by the verification of our predictions by means of protein docking of PR-proteins with the IgE antibody. This work also examines the conservation of IgE binding residues in allergen sequences and the evolution of different families of PR-Proteins. This study highlights that the other 11 classes of PR-proteins also possess the potential to be allergens.

Material and Methods

Sequence Retrieval and Allergenicity Prediction

The sequences for all PR-Proteins, both known and unknown allergens were retrieved from UniProt (<https://www.uniprot.org/>) (**Supplementary table 1**) [14]. 8 out of 19 proteins with established allergenicity status (**Supplementary table 2**) were also analysed to compare them with the remaining 11 proteins of unknown allergenicity status to infer any similarities and/or differences between their allergenicity potential. The allergenicity potential for each unknown allergen was predicted using the tools AlgPred (SVM and BLAST algorithms) [15], AllerTOP[16], AllergenFP [17] and PREAL^w (Motif-based and weighted average score algorithms) [18].

B-Cell Epitope Binding Predictions

The tool BepiPred v2.0 (<http://www.cbs.dtu.dk/services/BepiPred/>) was used to determine the B-cell epitope binding residues in all 19 classes of the PR-proteins. The threshold of epitope prediction was set to a score of 0.5 or a probability of 50% for a residue to be predicted to be an epitope. This threshold value was chosen to reduce the number of false positive (FP) predictions [19]. Two other tools, namely ABCpred (<https://webs.iitd.edu.in/raghava/abcpred/index.html>) [20] and Bcepred (<https://webs.iitd.edu.in/raghava/bcepred/index.html>) [21] were also used. The tool ABCpred uses artificial neural networks to predict the epitope. The threshold for prediction was set to 0.5 to reduce FP predictions. The epitope prediction sliding window was set to a 16 mer overlapping window. For the tool Bcepred, the predictions were made by using accessibility of amino acids as the physico-chemical property of the protein. Since the threshold for this tool varies in a range of [-3,3], therefore we used a more stringent value of 2 to get more true positives (TP). Once the predictions were made from all the tools, the sequences were compared with each other and a consensus sequence was derived

containing residues predicted as epitopes by all the three tools. Along with this, the numerical position of the residue has also been reported.

Protein Structures: Retrieval, Prediction, Validation and Visualisation

To study the allergen-antibody interactions of PR-proteins, the structures for PR-2 and the IgE molecule were retrieved from PDB (<https://www.rcsb.org/>) (**Supplementary table 3**) [22]. The structures for all other proteins were not available and hence, they were modelled initially using homology modelling approach on Swiss Model Portal (<https://swissmodel.expasy.org/>) using templates from PDB with an overall sequence identity and similarity of $\geq 50\%$. Only proteins PR-3, PR-9, PR-10, PR-11, PR-12 and PR-15 were modelled using this approach [23]. For other 12 proteins, the sequence similarity and identity was found to be less than the threshold and thus they were modelled using the RaptorX tool (<http://raptorx.uchicago.edu/StructPredV2/predict/>) which predicts the model using remote homology and threading approaches [24]. The protein structures modelled using either of the tools were validated by using the RAMPAGE tool (<http://mordred.bioc.cam.ac.uk/~rapper/rampage.php>). For the models to be accepted, a threshold of $\geq 80\%$ was selected for the residues to lie in the allowed and favourable regions of the Ramachandran plot (**Supplementary Figure 1**). The models for which the residues were found to fall below this threshold were rejected and the sequences were remodelled. After model validation, the structures were visualised using the PyMol software (<https://pymol.org/2/>) [25] in cartoon representation with protein being represented in purple and the consensus epitope binding residues highlighted in green.

Molecular Docking of PR-proteins with IgE and scatter plot formation

The protein structures for all the PR-proteins were docked to the IgE molecule using the tool ClusPro v2.0 (<https://cluspro.bu.edu/home.php>) in an antibody mode [12]. The results returned the list of clusters for different orientations of the allergen-antibody interactions along with the weighted interaction energy scores for each cluster. All the clusters were checked for the difference in their RMSD values in both bound and unbound states using structural alignment in PyMol. The cluster that had the highest negative interaction energy value associated with it and showed the least deviation in the RMSD values was selected for it being the most stable orientation. The allergen interacting chains of the IgE molecule and their interaction energies were noted down. After visualisation of these docking structures, a scatter plot was generated in R using the ggplot2 package [26] to highlight the correlation

of the count of predicted B-cell binding residues in an allergen (IgE binding sites) with the interaction energies for each stable orientation. Boxplots and density maps were also created for the datasets of residue lengths and the docking energies to determine the median residue length and the docking energy. These plots also aided for the exploration of the data by looking at the outlier values in each case and depicts the distribution of the data around mean and the skewness of data (**Supplementary Figure 4**).

Multiple Sequence Alignment and the tree of Evolution

The predicted B-cell epitope residues from all the protein sequence for all PR-families were extracted and were subjected to Multiple Sequence Alignment (MSA) in R using the MSA package [27]. The phylogenetic tree was constructed using the complete protein sequences in CLUSTAL X via neighbour joining method [28]. The output evolutionary tree was visualised in nexus format using the tool Dendroscope v3.5.9 [29,30].

Results

Allergenicity Prediction

The allergenicity of the 11 PR-Proteins with unknown allergenicity potential were predicted (**Table 1**) using 4 different tools, AlgPred was used with two different approaches, the support vector machine (SVM) based method which predicted that except PR-15 and PR-18, all other proteins are potential allergens, whereas for the approach based on BLAST, except PR-9, all other proteins were predicted to be as non-allergens. AllerTOP on the other hand predicted only PR-12 and PR-13 to be allergenic. The tool AllergenFP only predicted PR-19 as a probable-allergen whereas others were predicted to be as probable non-allergens. The fourth tool PREAL^w was also used with two different algorithms, the motif-based approach predicted PR-7, PR-9, PR-15, PR-16 and PR-18 as allergens and the rest as non-allergens, whereas, the second algorithm which employs the method of weighted average score for all the different parameters used by the tool predicted only PR-7 and PR-18 to be as allergens. Because each class of protein was found to be predicted as allergen by at least one tool, therefore the protein sequences for each of them along with the protein sequences of known allergens were subjected to B-cell epitope predictions.

Moreover, to figure out whether any significant similarities exist between 11 PR-proteins and already known allergens similarity search was performed on Allergen database (<http://www.allergenonline.org>). Seven of these PR-proteins with unknown allergenicity status show significant similarity to already known allergens, suggesting that there is a

possibility that these proteins can be a potential allergen; their similarity range from 48.7% to 84% (PR-6 is 48.7% similar to the allergen Sola 13.0101 from *Solanum lycopersicum*; PR-16 is 51.4% similar to the allergen Pin k 2.0101 from *Phleum pratense*; PR-11 is 61.4% similar to the allergen Per a 12.0101 from *Periplaneta americana*; PR-13 is 66.4% similar to the allergen Tri a 37.0101 from *Triticum aestivum*; PR-7 is 68.3 % similar to the allergen Cuc m 1.0101 from *Cucumis melo*; PR-18 is 71.6% similar to the allergen Phl p 4.0101 from *Phleum pratense*; PR-12 is 84% similar to Par h 1.0101 of *Dahlia merckii*) . However, no allergen similar to PR-9, PR-15, PR-17 and PR-19 were observed on Allergen database.

B-Cell Epitope binding prediction and their structural visualisation

The Protein sequences for all the 19 classes of PR-Proteins showed the presence of B-cell binding epitopes present in their sequences (**Table 2 and supplementary table 4**), thus confirming their allergenic behaviour. The predicted epitope binding residues for each protein from all the 3 different tools has been highlighted in red. Upon aligning the sequences with each other, only those residues that were predicted to be as epitopes by all the three tools were retained in the consensus sequence. The position of each residue predicted as epitope binder is also mentioned along with the sequence. The numerical values suggest that the residues are either present in a contiguous stretch or as discontinuous entities, i.e., the predicted amino acid residues either lie next to each other throughout the sequence in a long stretch or they are scattered throughout the protein sequences in smaller chunks and can either be located at the start, middle or towards the end of the sequence or can either be a single amino acid residue such as for PR-1, the epitope binding residues lies in the range from residue 27th – 35th whereas for PR-2, the epitope binding residues lie scattered throughout the sequence, i.e., they lie within the range of 38th – 44th residue (beginning of the protein sequence or the N-terminal), 229th – 235th residue (middle of the protein sequence) and 332nd – 334th residue (end of the protein sequence or the C-terminal). Along with these chunks, there are single amino acid residues at the 150th, 294th and 302th residue position. This suggests that depending upon the accessibility of the amino acids to be on the surface of the allergen protein, the amino acids have been predicted to be a B-cell binding epitope. Apart from this, in case of PR-6, out of 223, 129 residues have been predicted to be epitope binders suggesting that around 57.84% of residues can potentially interact with the Immunoglobulins and can trigger a strong allergenic response. It is also conspicuously visible that the epitope binding sites are

separated by small number of residues, thus indicating the continuous distribution of these residues on the protein surface. **Figure 2** and **Supplementary Figure 1** represents the 3D structure of all the PR-Proteins as a cartoon model. The protein has been coloured in purple with epitope binding residues highlighted in green.

Protein Modelling and Structure Validation

Since the structure for only PR-2 was retrieved from PDB, therefore the structures for other proteins were modelled. The proteins PR-3, PR-9, PR-10, PR-11, PR-12 and PR-15 were modelled using Swiss Model Portal (SMP). To model these proteins, the best template was chosen based upon the sequence similarity, sequence identity and the query coverage returned for each protein sequence after alignment with different templates taken from PDB. Out of various models built, the model with highest GMQE (Global Model Quality Estimation) score was chosen. The GMQE score represents the reliability of the model. As the score lies between 0 and 1, therefore higher the score better is the estimation. For each of the proteins modelled, the GMQE score was closer to 1. Thus, suggesting that all the models were highly reliable. Also, the QMEAN Z-score also provides the quality estimation, both on global and on local scales and indicates how close the model is to its' native folds. The QMEAN Z-score is a measure of 4 different parameters, namely, the interaction potential between the C_β atoms, the interaction potential between all atoms of the protein residues, the solvation potential of the protein and the torsion potential. A QMEAN Z-score closer to zero or numbers greater than -4.0 are regarded as good because the closer the value to 0, the closer the model is to the experimentally determined structures. The QMEAN Z-score for all of the 6 proteins modelled was found to be closer to 0 and higher than -4.0, thus suggesting that our models were closer to its native structural folds that are present in the PDB templates taken for their modelling (**Table 3**).

The other tool RaptorX, uses remote homology and threading approaches for protein modelling and thus, the remaining 12 proteins were modelled using this tool. It searches for the templates that shares the homology with specific folds of the proteins depending upon its amino acid composition. These folds are defined by the Structural Classification of Proteins (SCOP) according to which superfamily the protein belongs, such as all alpha, all beta, both alpha and beta etc. Also, the quality estimation in this tool is done by using the p-values comparison, the global distance test score (both normalised and unnormalized;

uGDT, GDT), the alignment score and the uSeqID/SeqID score which represents the number of identical residues found in the alignment of the query sequence with various other proteins that are used as templates to model the 3D structure of the query protein with the correct fold. Since, the models generated by the tool are ranked based on these values, therefore the p-value indicates the relative quality of the models. Therefore, it is a rule of thumb that the lower the p-value for a model is, the better is the predicted structure. But since our protein sequences were of varying lengths, therefore the uGDT and GDT score values were also considered to choose the absolute quality model. For the protein sequences of length > 100, a uGDT score of > 50 and for protein sequences of length < 100, a GDT score of > 50 was considered as a measure of the model. The alignment score also makes an important quality check measure. The score usually lies between 0 and the length of the query sequence. Therefore, the higher the alignment score, the better is the model. A similar rule follows for the uSeqID score, the higher the score is, the model has a higher chance to have the correct fold (**Table 4**). But as an exception, the structure of PR-13 (both domain 1 and 2) as predicted did not have the expected uGDT/GDT or uSeqID/SeqID scores. But since the p-value and the alignment scores were acceptable, therefore the model was retained for the docking analysis.

When all the proteins were modelled, their model quality was validated using the Ramachandran Plot analysis (**Supplementary Figure 2**). For all the proteins, the threshold for the residues to be accepted as a good and reliable model was set to $\geq 80\%$. It was found that except for the proteins PR-6 and PR-13, the percent residues that lie in the favourable regions of the Ramachandran plot was more than 90% with very less residues in the allowed and outlier regions. This indicated that our modelled proteins have the correct (ϕ, ψ) pair of angles between the residues. Whereas for the proteins PR-6 and PR-13, the percent of residues in favoured regions was 88.7% and 84.1% with 2.7% and 4.5% residues falling in the outlier regions respectively. It was only 6 residues in each case that did not have the correct (ϕ, ψ) pair of angles but since a high number of modelled residues were above the threshold for them to be accepted as a good model, the models for both the proteins were retained to study their interactions with the IgE molecules.

Visualising Allergen-Antibody Interactions via Molecular Docking and the correlation of interacting residue lengths with docking energies

The docking results as returned by the tool ClusPro were visualised in PyMol. The allergen has been represented in cartoon form (with predicted epitope binding residues highlighted in green) and the IgE molecule has been represented in surface view (allergen interacting chains are highlighted in orange). The docking images depict the interactions between the two proteins, the IgE antibody and the PR-allergen (red dashed lines) (**Figure 2a, 2b, 2c and , 2d and supplementary figure 3**). These lines represent the bond length of the hydrogen bonds formed between the allergen and the antibody. The IgE molecule has 12 chains (A to L), thus it was found that each allergen molecule interacts with only a few chains of the antibody molecule. The interaction energy along with the respective interacting chains of the IgE molecule with each allergen was recorded (**Table 5a and 5b**). It was observed that for the PR-5 family of proteins, the interaction energy was more positive (-366.1 kJ/mol) in comparison to other interactions. On the other hand, for PR-6 family of proteins, the interaction energy was found to be the most negative (-545.0 kJ/mol). It was also observed that for the proteins PR-9, PR-11, PR-14 and PR-18 that they interact with only 3 chains (chains H, I and L) of the IgE molecule but there is a vast difference between their interaction energies with the antibody, with PR-11 = -383.8 kJ/mol and PR-14 = -486.2 kJ/mol of energy with the docking energy for PR-9 and PR-18 lying between this range. This could possibly indicate that this difference of interaction energy between these proteins is because of differences in their respective protein folds and the orientation in which they interact with the IgE chains. As the interactions involve strong inter and intra hydrogen bonding, therefore it can be anticipated that the bond strength might have a role to play that can contribute this difference between the interaction energies.

A Pearson's correlation coefficient was also calculated between the docking interaction energies (in kJ/mol) and the IgE binding residue counts (number of predicted epitopes binding amino acid residues). The correlation coefficient $r = 0.4851$ at a p -value = 0.0352, CI = (0.03970523 – 0.76974762) was found to be statistically significant with a scatter plot depicting a weak positive linear correlation between the two variables (**Figure 3**). It is conspicuously visible in the plot that the proteins PR-2, PR-5, PR-6, PR-9, PR-10, PR-11, PR-13, PR-14, PR-15, PR-16 and PR-17 do not lie very closer to the line of best fit and deviate from the trend represented by the other 8 PR-protein families. This could be the reason for an overall weak positive correlation between the two variables. The boxplots and the density plots obtained for the data sets also suggests that the data is partially skewed

and thus there might be a slight deviation in the trend of correlation between the two variables (**Supplementary Figure 2**).

Multiple Sequence Alignment and the tree of evolution

The Multiple Sequence Alignment (MSA) of the B-cell epitope binding residues consensus sequence for all the PR-Proteins was carried out which suggested that the antibody interacting residues in all the classes of proteins do not show a high degree of conservation (**Figure 4**). The residue positions marked in asterisk (*) shows that the residues at these positions are either conserved with a rate of more than 25% or shows a similar substitution. The residues marked in blue are conserved across different PR-families, such as the residue lysine (K) is conserved in 6 families (PR-3, PR-10, PR-11, PR-12, PR-15 and PR-16) as marked with a red arrow. Similarly, at the position marked with a green arrow shows that the residues at these positions show similar residue substitution.

Since, the MSA output suggested a low rate of epitope binding residue conservation throughout all the classes of proteins, therefore a phylogenetic tree was constructed using the complete protein sequences to understand the inter-relationship of these families over a complete sequence length. The phylogenetic tree obtained represents the evolutionary distances for each class of proteins with their immediate common ancestor, with other PR-protein families and with the root of the tree, which represents the common ancestor for all the PR-protein families. The scale at the bottom of the tree represents the percentage of genetic variation. The scale value here is 0.10, which implies that a rate of 10% genetic variation has been observed in the protein sequences. The distance on each branch represents the distance of that family of PR-proteins from their common ancestor from which the two families diverged during the course of evolution. The tree of PR-proteins is majorly divided into three clades where each clade has the most similar protein families clustered together. For example, in clade 1, the PR-15 and PR-16 proteins are clustered together suggesting a high degree of similarity between them. These are then clustered with the PR-9 proteins suggesting that there is a high sequence similarity between the two. A similar explanation is true for the proteins PR-1, PR-17 and PR-11 that have been clustered together and these two small clusters makes the first clad. Thus, we can say that the proteins PR-15, PR-16, PR-9, PR-1, PR-17 and PR-11 share a significant sequence similarity and have evolved similarly in time. This holds true for the proteins in other clads as well. The

branch lengths also indicate the distance between the two families of the proteins, i.e., the sum of the branch lengths when going from one family to another is the distance between the families or in other words, it is the genetic distance between the two sequences. For instance which since the distance between the families PR-15 and PR-9 is less than the distance between the families PR-15 and PR-1, therefore PR-15 is clustered with PR-9 (Figure 5).

Discussion

Approximately, 30% of the population of the developed countries suffer from common health problems such as asthma, rhinitis, eczema and other diseases, which has been classified to be as allergic diseases. PR-Proteins is a class of proteins that are synthesized by the plants and is present in various plant parts, such as roots, stem, leaves, pollen grains and fruits [31,32]. This makes these allergens to fall under the categories of both airborne and food allergens. Upon inhalation of pollen grains in air or the ingestion of different plant parts as food, these proteins enter the human body and trigger an allergenic response such as swelling of lips, watery eyes, nausea etc. Therefore, it is important to understand the allergenic potential of these proteins. We performed an *in-silico* analysis on all the 19 classes of PR-Proteins, thus establishing that all the classes have the potential to trigger an allergenic response. Out of 19 different families, PR-1, PR-2, PR-3, PR-4, PR-5, PR-8, PR-10 and PR-14 were already defined as allergens. Therefore, we predicted the allergenicity for the other 11 proteins. We established from our findings that all the PR-Protein families have the potential to act as allergens and can trigger the allergenic response. It is thus important to identify their presence in different plant parts used as food to prevent creating life threatening conditions such as anaphylaxis and rhino conjunctivitis.

The allergy associated symptoms appear in a patient when these allergens enter the human body and interact with the antibody molecules and hence, it is also important to decipher how these allergen molecules interact with the antibodies (IgA, IgD, IgE, IgG and IgM) being synthesized in our bodies [9]. The B-cell epitope binding predictions thus provides evidence about the residues that interacts with the antibody molecules. From our findings about the 19 PR-Protein families, we concluded that the residues that interact with the antibody molecules usually lie on the surface of the proteins and are solvent accessible. Thus, these residues are hydrophilic and interact with the water molecules present as a

solvent in the interacting system. It is also generally hypothesised that if an allergen contains more epitope binding residues, the stronger are the interactions with the antibodies and severe is the allergenic response. The severity of the response also depends upon the concentration of the allergen. For example, banana contains various PR-proteins and thus forms an ideal allergen study system. If the banana fruit has a very high concentration of these proteins, then upon its consumption severe allergenic responses can be triggered. Apart from the concentration, these PR-Proteins can also cross-react and can worsen the symptoms. One such case is the cross-reactivity between the PR-3 and PR-4 allergens that can cause latex-fruit syndrome. Thus, we suggest, a prior knowledge of the B-cell epitope binding residues present in these proteins would be advantageous for clinicians to treat the symptoms in a more comprehensive way.

Since, these allergenic responses are triggered upon molecular interactions of the allergen-antibody system, therefore, it is of utmost importance to understand the structure of allergens. These allergen molecules need to have the correct 3D shape with the correct fold to be functional and elicit a response. The correct folding of these allergens depends upon 4 different factors, namely, the hydrophobic effect in which the hydrophobic chains of the protein are buried to form a core and the hydrophilic residues constituting the surface of the protein molecule; Van der Waals forces that gives rise to the attraction and repulsion forces due to the formation of transient dipoles during the protein folding process; Hydrogen bonding with the water molecules (solvation potential) and the thermodynamic factors (enthalpy and entropy). The thermodynamic stability of the protein depends upon the change in Gibb's free energy (ΔG) during the folding and unfolding of the proteins. The change in ΔG rely upon the changes in enthalpy (ΔH) and entropy (ΔS). Thus, protein modelling and the criteria to accept the 3D model of the protein should be very stringent. Whether our modelled proteins had the correct fold or not was thus verified using the Ramachandran plot. When a protein folds to attain a functional conformation the bond angles Phi (ϕ) and Psi (ψ) provides the correct degree of rotation of the residues in the intersecting planes to avoid steric clashes. Our protein structure validation suggested that all our modelled proteins had the correct fold with the correct (ϕ, ψ) pair of angles and thus are functional to be as allergens and interact with the antibody molecules.

Molecular docking of the allergen molecules with the IgE antibody has provided an additional evidence to support our hypothesis that all PR-Protein families are potential allergens. The allergens interact with the IgE molecules in a specific orientation which is energetically most favourable. The more negative the interaction energy, the more stable is the interaction. But this is not always the case as proteins interact dynamically with other molecules. That is, the proteins constantly modify their shapes to get accommodated in the most favourable orientation to provide stability to the interaction system as a whole. Thus, the difference between the root mean square deviation (RMSD) values of the proteins in both bound and unbound states qualifies to be a good quality check. We found that the RMSD values for all the 19 proteins did not show a significant difference between their bound and unbound states. The RMSD values for 10 proteins was found to be $< 1\text{\AA}$ and for the other 9 proteins, they were $< 2\text{\AA}$ (data not shown). Therefore, it was certain that our docking results were in accordance with the basic principles of protein folding and molecular protein-protein interactions. The docking interaction energies reported in the results section is a sum total of the protein-protein interaction stability factors mentioned above and another quantity known as the statistical or knowledge potential or DARS (Decoys As the Reference State) potential. The DARS potential finds the near native structures of the docking proteins, calculates the values for other potentials and generates a pair-wise structure based intermolecular potential for the interacting atoms. This also provides the dissolution contribution in the form of ΔG which is created upon the initial removal of the water molecules from the docking system. Thus, it aids in visualising the dynamic behaviour of the interacting biomolecular systems [33].

Though the docking interaction energies varies mainly due to Van der Waals forces, electrostatic potentials, hydrogen bonding, hydrophobic effects and other factors contributing to protein stability and interactions, we hypothesized that there might be a relationship between the IgE binding residue lengths and the docking energies. Our hypothesis directed us to establish that there is a weak positive correlation between the two quantifiable variables. But as evident from the plot that all proteins were not found to lie closer to the line of best fit, therefore, we suggest that for these proteins, apart from the IgE binding residue lengths there could be other factors that might have contributed towards this deviation. Factors such as disorder in the predicted protein structures, i.e., whether the residues have been modelled to lie in the correct fold of the protein backbone

or not; thermodynamic stability of the folded protein, i.e., the ability to stay in its native functional form; the presence of more loops in the predicted structures than helices and sheets; solvent exposure of the hydrophobic chains and the delocalisation of the π -electron clouds of the aromatic amino acids might have played a significant role. Thus, it is important to decipher the crystal structure of these proteins either by NMR spectroscopy or X-ray crystallography to understand how these factors might play a role in the correct protein folding and unfolding. Therefore, we strongly suggest that the energetics of these proteins should be studied to understand the mechanism behind their ability to attain specific structural folds and to become functional.

The MSA of the IgE binding regions was also done to identify highly conserved residues among all the PR-Protein families. But the rate of conservation was not found to be high and thus no relationship between the allergenicity potential of different families of the PR-Proteins can be established with certainty. But it can be stated certainly that all the different PR-families does have the potential to trigger an allergic response, but its severity cannot be predicted without correlating the impact with clinical data. Therefore, to understand the impact of the non-conservation of these residues, some clinical allergenicity tests (skin prick, blood test, patch test and food elimination tests) should be carried out. This will throw some light upon the allergenic potential of different allergens from different families and would also help to identify various cross-reactivity associated allergic symptoms.

Since, no conservation pattern was observed in the IgE binding residues, therefore, by using the entire protein sequences a phylogenetic tree was constructed to understand the evolutionary relationship between different PR families. We hypothesise that the sequences might have diverged in time either due to some evolutionary forces such as selection and mutation or maybe due to their different functions apart from being allergens. For example, the events of speciation might have modified the sequences either at the protein level or the nucleotide level to evolve them to perform different functions, such as endochitinase, proteinase inhibitor, peroxidase, oxidoreductase etc [3]. As these proteins have various other roles in plant development, therefore, it can also be attributed that these proteins might have evolved to take up some specific functions during different developmental stages, such as flower development, pollen germination etc. Another explanation could be a high degree of recombination between the homologous sequences

that might have led to a high degree of divergence or less similarity between the sequences. Since, these proteins are also known as the stress proteins, therefore, the evolutionary forces might have diverged them to function specifically under different situations of both biotic and abiotic stresses, i.e., the protein sequences might have diverged as a consequence of co-evolution with the pathogens and herbivores to provide a fine line of defence against them and due to continuous exposure to several abiotic stresses, such as salt stress, heavy metal stress or drought stress etc. Hence, we conclude by stating that the sequences of all PR-families do not show a very high degree of residue conservation and thus have been evolved to have very many different functions. Therefore, we recommend *in vitro* investigation of these proteins and correlating these findings with the severity of the clinical symptoms.

Concluding Remarks

Allergenicity is a major cause of concern as a huge proportion of population suffers from one or the other allergenic disease for which many times the cause remains undetected. Hence, it is important to understand about the presence of allergens in the food products that are consumed on everyday bases. Our *in-silico* analysis of different families of PR-Proteins suggested that these proteins can trigger an allergenic response and can even trigger a cross-reactivity response. Henceforth, we strongly support that some prior knowledge about B-cell epitope binding residues and the interactions between allergens and antibodies could provide an important insight to understand various allergenic responses invoked by the different families of PR-Proteins. This can help clinicians to provide a tailor-made dose of various drugs available to treat different symptoms triggered by different allergens. We also conclude that because proteins are dynamic in nature, therefore it will be interesting to see how they interact with other antibody molecules and how does the different forces operating during the protein-protein interactions stabilize them to attain a particular orientation with the most stabilising interaction energy.

Acknowledgement

The authors acknowledge Science and Engineering Research Board (SERB), DST, Ministry of Science and technology, Govt. of India, New Delhi, India, for the financial supports given to AS and IKS [SB/FT/LS-188/2012 (AS); ECR/2017/002478 (AS); SB/YS/LS-59/2013 (IKS)].

Authors Contribution

AS and IKS conceived the idea, planned the work and wrote the paper. RA and AK performed the *in silico* work, analysed data and designed the figures. All authors discussed the results and contributed to the final manuscript.

References

- [1] A. Stintzi *et al.*, "Plant 'pathogenesis-related' proteins and their role in defense against pathogens," *Biochimie*, vol. 75, no. 8, pp. 687–706, 1993.
- [2] T. Midoro-Horiuti, E. G. Brooks, and R. M. Goldblum, "Pathogenesis-related proteins of plants as allergens," *Annals of Allergy, Asthma and Immunology*, vol. 87, no. 4. pp. 261–271, 2001.
- [3] M. Sinha *et al.*, "Current overview of allergens of plant pathogenesis related protein families," *The Scientific World Journal*, vol. 2014. 2014.
- [4] M. Gajhede *et al.*, "X-ray and NMR structure of Bet v 1, the origin of birch pollen allergy," *Nat. Struct. Biol.*, vol. 3, no. 12, pp. 1040–1045, 1996.
- [5] H. Breiteneder and E. N. C. Mills, "Plant food allergens - Structural and functional aspects of allergenicity," *Biotechnology Advances*, vol. 23, no. 6. pp. 395–399, 2005.
- [6] M. Mayoral, H. Calderón Moreno, R. Cano, and M. Lombardero, "Allergic rhinoconjunctivitis caused by Cannabis sativa pollen," *J. Investig. Allergol. Clin. Immunol.*, vol. 18, no. 1, pp. 73–74, 2008.
- [7] A. Leitner *et al.*, "Allergens in pepper and paprika: Immunologic investigation of the celery-birch-mugwort-spice syndrome," *Allergy Eur. J. Allergy Clin. Immunol.*, vol. 53, no. 1, pp. 36–41, 1998.
- [8] S. Wagner and H. Breiteneder, "The latex-fruit syndrome," *Biochem. Soc. Trans.*, vol. 30, no. 5, p. A109.3-A109, 2015.
- [9] A. D. Dulaney, "Essentials of Immunology," *Am. J. Public Health*, vol. 15, no. 11, pp. 1000–1000, 2008.
- [10] K. Senger and A. A. Zarrin, "Structure and Function of IgE," in *Encyclopedia of Immunobiology*, vol. 2, 2016, pp. 31–39.
- [11] A. D. Baxevanis and B. F. F. Ouellette, *Bioinformatics : a practical guide to the analysis of genes and proteins*, no. 43. 2001.
- [12] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: A fully automated algorithm for protein-protein docking," *Nucleic Acids Res.*, 2004.
- [13] D. Kozakov *et al.*, "How good is automated protein docking?," *Proteins Struct. Funct. Bioinforma.*, 2013.
- [14] EMBL, SIB Swiss Institute of Bioinformatics, and Protein Information Resource (PIR), "UniProt," in *Nucleic acids research*, 2013, pp. 41: D43-D47.
- [15] S. Saha and G. P. S. Raghava, "AlgPred: Prediction of allergenic proteins and mapping of IgE epitopes," *Nucleic Acids Res.*, vol. 34, no. WEB. SERV. ISS., 2006.
- [16] I. Dimitrov, I. Bangov, D. R. Flower, and I. Doytchinova, "AllerTOP v.2-a server for in silico prediction of allergens," *J. Mol. Model.*, vol. 20, no. 6, p. 2278, 2014.
- [17] I. Dimitrov, L. Naneva, I. Doytchinova, and I. Bangov, "AllergenFP: Allergenicity prediction by descriptor fingerprints," *Bioinformatics*, vol. 30, no. 6, pp. 846–851, 2014.
- [18] J. Li, J. Wang, and J. Li, "Improving Allergen Prediction in Main Crops Using a

- Weighted Integrative Method," *Interdiscip. Sci. Comput. Life Sci.*, 2017.
- [19] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "BepiPred-2.0: Improving sequence-based B-cell epitope prediction using conformational epitopes," *Nucleic Acids Res.*, 2017.
 - [20] S. Saha and G. P. S. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins Struct. Funct. Genet.*, 2006.
 - [21] S. Saha and G. P. S. Raghava, "BcePred: Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties," 2010.
 - [22] Protein Data Bank, "RCSB PDB: Homepage," *RCSB PDB*, 2019. .
 - [23] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede, "The SWISS-MODEL Repository and associated resources," *Nucleic Acids Res.*, 2009.
 - [24] M. Källberg *et al.*, "Template-based protein structure modeling using the RaptorX web server," *Nat. Protoc.*, 2012.
 - [25] W. L. DeLano, "The PyMOL Molecular Graphics System, Version 1.1," *Schrödinger LLC*, 2002.
 - [26] H. Wickham, "ggplot2," *Wiley Interdiscip. Rev. Comput. Stat.*, 2011.
 - [27] U. Bodenhofer, E. Bonatesta, C. Horejš-Kainrath, and S. Hochreiter, "Msa: An R package for multiple sequence alignment," *Bioinformatics*, 2015.
 - [28] M. A. Larkin *et al.*, "Clustal W and Clustal X version 2.0," *Bioinformatics*, 2007.
 - [29] I. Ebersberger, A. Von Haeseler, and H. A. Schmidt, "Phylogeny Reconstruction," in *Bioinformatics-From Genomes to Therapies*, 2008.
 - [30] D. H. Huson and C. Scornavacca, "Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks," *Syst. Biol.*, vol. 61, no. 6, pp. 1061–1067, 2012.
 - [31] A. Mari, E. Scala, P. Palazzo, S. Ridolfi, D. Zennaro, and G. Carabella, "Bioinformatics applied to allergy: Allergen databases, from collecting sequence information to data integration. The Allergome platform as a model," *Cell. Immunol.*, vol. 244, no. 2, pp. 97–100, 2006.
 - [32] J.-J. Liu and A. K. M. Ekramoddoullah, "The family 10 of plant pathogenesis-related proteins: Their structure, regulation, and function in response to biotic and abiotic stresses," *Physiol. Mol. Plant Pathol.*, vol. 68, no. 1–3, pp. 3–13, 2006.
 - [33] R. Brenke *et al.*, "Application of asymmetric statistical potentials to antibody-protein docking," *Bioinformatics*, 2012.

Figure legends:

Figure 1. PR-proteins as allergen. The PR-proteins are expressed in different plant parts and also upon exposure to biotic and abiotic stresses and are accumulated into fruits, seed and pollen grains; when these proteins are ingested or inhaled, they interact with the IgE molecules. The IgEs are structurally complex and consist of 12 different protein chains that can interact with different residues of the allergens, triggering an immune response via antibody-allergen interaction, finally leading to specific allergenic response and development of symptoms such as oral allergy syndrome, asthma, anaphylaxis, conjunctivitis, Rhinitis, nausea, hypertension etc.

Figure 2. The 3D structure of different classes of PR-proteins and; molecular docking between PR-proteins and the IgE molecules. 2(a) PR-6, 7 & 8 ; **2(b)** PR-9, 11 & 13 ; **2 (c)** PR-15, 16, & 17; **2 (d)** PR-18 & 19 . The structures have been represented in cartoon form in purple-blue colour with the epitope binding residues highlighted in green. The antibody (orange) has been represented in surface view. The image in the center represents an overview of the docking between PR-proteins and IgE molecule and the image on the right depicts the zoom in view of the interactions (in red). The numbers on the dashed lines (interactions) represents the bond lengths between the interacting atoms.

Figure 3. Correlation between IgE binding residue count and docking energy. The dotted line represents the line of best fit for a linear model between the two variables. Each dot represents the corresponding PR-Protein that lies inside or outside of the 95% confidence interval (CI) (in blue around the dotted line).

Figure 4. Multiple Sequence Alignment of predicted B-cell epitope binding residues. The blue boxes show that the residues at each of the marked locations shows a high conservation pattern whereas the pink boxes highlight the residues at the locations that show a conserved substitution. The asterisk marked under the highlighted boxes suggests a conservation pattern of these residues.

Figure 5. Phylogenetic Tree. PR-Proteins have been clustered here using the neighbour-joining method. Each terminal node represents one family of PR-Proteins and internal nodes represents the immediate ancestor of the clustered protein family. The distances over the edges represents the distance of each protein family from its immediate ancestor.

Table1: Allergenicity predictions for PR-proteins of unknown allergenicity status

Family	Source	Protein	AlgPred		AllerTOP	AllergenFP	PREAL ^w	
			SVM-method	BLAST-method			Motif-based method	Weighted-average score method
PR-6	<i>Solanum lycopersicum</i>	Proteinase Inhibitor	Potential Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Non-Allergen	Non-Allergen
PR-7	<i>Solanum lycopersicum</i>	Pathogenesis related protein P69G	Potential Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Allergen	Allergen
PR-9	<i>Nicotiana tabacum</i>	Lignin forming peroxidase precursor (EC 1.11.1.7)	Potential Allergen	Allergen	Probable Non-Allergen	Probable Non-Allergen	Allergen	Non-Allergen
PR-11	<i>Nicotiana tabacum</i>	Chitinase, class V	Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Non-Allergen	Non-Allergen
PR-12	<i>Dahlia merckii</i>	Defensin-like protein 1	Allergen	Non-Allergen	Probable Allergen	Probable Non-Allergen	Non-Allergen	Non-Allergen
PR-13	<i>Arabidopsis Thaliana</i>	Thionin 2.1	Allergen	Non-Allergen	Probable Allergen	Probable Non-Allergen	Non-Allergen	Non-Allergen
PR-15	<i>Hordeum Vulgare</i>	Oxalate oxidase	Non-Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Allergen	Non-Allergen
PR-16	<i>Hordeum vulgare</i>	Oxalate oxidase like or germin-like protein	Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Allergen	Non-Allergen
PR-17	<i>Nicotiana Tabacum</i>	NtPRp27	Potential Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Non-Allergen	Non-Allergen
PR-18	<i>Helianthus Annuus</i>	Carbohydrate oxidase	Non-Allergen	Non-Allergen	Probable Non-Allergen	Probable Non-Allergen	Allergen	Allergen
PR-19	<i>Pinus sylvestris</i>	Antimicrobial peptide	Allergen	Non-Allergen	Probable Non-Allergen	Probable Allergen	Non-Allergen	Non-Allergen

Table 2: B-cell epitope binding residues predictions of 11 PR-proteins with unknown allergenicity status

Family	Source	Protein	B-cell Epitpe Prediction				Residue Position
			BepiPred	ABCpred	Bcepred	Consensus	
PR-6	Tomato	Proteinase Inhibitor	MAIYKVALLLFGMI LLASDFEHAKACTKE CDTRIDFGICPLETK RVEGLCTNCCAGKK GCKYFSKDGTYICDG ESEWVSEKNNNLKK ACTKECDTRIDFGIC PLETEKRVGLCTNC CAGKKGCKYFSKDG TYICDGESEWVSEKD NNLEKDKTECDTRI DFGICPLETKRVKG LCTNCCAGKKGCKY FSADGTYICDGESE WVSEGENDLQKSN VAIS	MAIYKVALLLFGMI LLASDFEHAKACTKE CDTRIDFGICPLETK RVEGLCTNCCAGKK GCKYFSKDGTYICDG ESEWVSEKNNNLKK ACTKECDTRIDFGIC PLETEKRVGLCTNC CAGKKGCKYFSKDG TYICDGESEWVSEKD NNLEKDKTECDTRI DFGICPLETKRVKG LCTNCCAGKKGCKY FSADGTYICDGESE WVSEGENDLQKSN VAIS	MAIYKVALLLFGMILL ASDFEHAKACTKECDT RIDFGICPLETKRVEGL CTNCCAGKKGCKYFSK DGTYICDGESEWVSEK NNNLKACTKECDTRI DFGICPLETKRVEGLC TNCCAGKKGCKYFSK GTICDGESEWVSEKD NNLEKDKTECDTRIDF GICPLETKRVKGLCTN CCAGKKGCKYFSADGT YICDGESEWVSEGEN LQKSNVAIS	-----EHAK- -TKECDTRI----- PLETEKRVG----- KKGCKYFSK----- SEWVSEKNNNLKACT KECDTRI----- PLETEKRVG-- ----- GKKGCKYFSK----- SEWVSEKNNLEKDKT KECDTRI----- PLETEKRVKGL----- GKKGCKYF----- SEGENDLQKSNV---	22-25, 28-35, 41-50, 59-68, 77-99, 105- 114, 122-132, 141- 162, 169-179, 186- 193, 209-220
PR-7	Tomato	PRp69G	MEKKHGFVSAQKQ RILSLHTHTPSFLGL QQNMGLWKDSNF GEGVIIGVLDTGILPD HPSFSDVGMPPPPA KWKGVCEPNFTNKC NNKLIGARSYKLG GSPIDDDGHGTHTA STAAGAFVKGANVY ENANGTAVGVAPLA HIAIYKCVNSVGCSD SDILAAMDSAIDDG VDILSISIGSLRPLYD ESIALGAYSATQRGIL VSCSAGNNGPSPAS VDNSAPWILTVGAS TLDRKIKATAKLGNG EEFEGESAYRPKISNS TFFTLFDAAKNAKDP SETPYCRRLTDP IRGKIVLC LAFGGVANVDKQAV NVDKQAVKDAAGG VGMIVINPSQYGV KSADAHVLPALVVS AADGSKIRAYTNSIS N	MEKKHGFVSAQKQ RILSLHTHTPSFLGL QQNMGLWKDSNF GEGVIIGVLDTGILPD HPSFSDVGMPPPPA KWKGVCEPNFTNKC NNKLIGARSYKLG GSPIDDDGHGTHTA STAAGAFVKGANVY ENANGTAVGVAPLA HIAIYKCVNSVGCSD SDILAAMDSAIDDG VDILSISIGSLRPLYD ESIALGAYSATQRGIL VSCSAGNNGPSPAS VDNSAPWILTVGAS TLDRKIKATAKLGNG EEFEGESAYRPKISNS TFFTLFDAAKNAKDP SETPYCRRLTDP IRGKIVLC LAFGGVANVDKQAV NVDKQAVKDAAGG VGMIVINPSQYGV KSADAHVLPALVVS AADGSKIRAYTNSIS N	MEKKHGFVSAQKQRL SLHTHTPSFLGLQQN MGLWKDSNFGEGVIIG VLDTGILPDHPSFSDVG MPPPPAKWKGVCEPN FTNKCNNKLIGARSYK GNGSPIDDDGHGTHT ASTAAGAFVKGANVY NANGTAVGVAPLAHIA YKVCNSVGCSDSILAI AMDASIDGVDILSISI GSLRPLYDESIALGAY SATQRGILVSCSAGN GPSPASVDNSAPWILT VGASTLDRKIKATAKLG NGEEFEGESAYRPKIS STFTFLFDAAKNAKDPS ETPYCRRLTDPAIRG KIVLC LAFGGVANVDKQAV KDAGGVGMIVINPSQY GVTKSADAHVLPALV SAADGSKIRAYTN SISN	---HGFVSAQKQ-ILS- HTTHTPS----- KDSNFG----- -----MPPPPAKWKG-- EPNFTNKCNNKL--- RSYKLG----- ----- NVYENANGT----- ----- -----LRPLYDE----- ----- -----RKIKAT--- NGEE----- AYRPKISNST---- DAAKNAKDPSETPYCR RGSL----- DKGQ----- ----- ---	5-13, 15-17, 19-25, 37-43, 66-75, 78-89, 93-99, 125-133, 182- 188, 235-240, 244- 248, 254-263, 269- 288, 310-313
PR-9	Tobacco	Lignin forming Peroxidase precursor (1.11.1.7)	MSFLRFVGAILFLVAI FGASNAQLSATFYD TTCPNVTSIVRGVM DQRQRTDARAGAKI IRLHFHDFVNGCD GSILLDTDGTQTEKD APANVGAGGFIDV DIKTALENVCPGVVS CADILALASEIGVLLA KGPSWQVLFGRKDS LTANRSGANSIDIPSP FETLAVMIPQFTNK GMDLTDLVALSGAH TFGRARCGTFEQRFL NFNGSGNPDLTVD TFLQTLQGICPQGG NNGNTFTNLDISTP NDFDNDYFTNLQSN	MSFLRFVGAILFLVAI FGASNAQLSATFYD TTCPNVTSIVRGVM DQRQRTDARAGAKI IRLHFHDFVNGCD GSILLDTDGTQTEKD APANVGAGGFIDV DIKTALENVCPGVVS CADILALASEIGVLLA KGPSWQVLFGRKDS LTANRSGANSIDIPSP FETLAVMIPQFTNK GMDLTDLVALSGAH TFGRARCGTFEQRFL NFNGSGNPDLTVD TFLQTLQGICPQGG NNGNTFTNLDISTP NDFDNDYFTNLQSN	MSFLRFVGAILFLVAIF GASNAQLSATFYDTTC PNVTSIVRGVMDQRQ RTDARAGAKIIRLHFH CFVNGCDGSILLDTDG TQTEKDAPANVGAGG FDIVDDIKTALENVCPG VVSADILALASEIGVV LAKGPSWQVLFGRKDS LTANRSGANSIDIPSP TLAVMIPQFTNKGM LTDLVALSGAHTFGRA RCGTFEQRFLNFNGSG NPDLTVDATFLQTLQGI CPQGGNNGNTFTNLDI STPNDFDNDYFTNLQS NQGLLQTDQELFSTSG SATIAIVNRYAGSQTF	----- -----RQRTD-- ----- DTDGTQTEKDAPAN--- ----- ----- GRKDSLANTANRSGANS -----NKG----- ----- ----- DISTPNDFDNDYFTNL QSNQG----- -----GSQTQFF----- -----TNGQI-----	47-51, 78-91, 142- 157, 174-176, 242- 262, 287-293, 312- 316

			QGLLQTDQELFSTSG SATIAIVNRYAGSQT QFFDDFVSSMIKLG NISPLTGTNGQIRTD CKRVN	QGLLQTDQELFSTSG SATIAIVNRYAGSQT QFFDDFVSSMIKLG NISPLTGTNGQIRTD CKRVN	FDDFVSSMIKLGNISPL TGTNGQIRTDCKRVN		
PR-11	Tobacco	Chitinase Class V	MANSVTLFAIIFSCFL LQQLVCTNSQNVKG GYWFKDSGLALNNI DSTLFTHLFCFADL NPQLNQLIISPENQD SFRQFTSTVQRKNPS VKTFLSIAGGRANST AYGIMARQPNRSRKS FIDSSIRLARQLGFH GLDLWEYPLSAAD MTNLGTLNEWRTA INTEARNSGRAALL TAAVSNSPRVNGLN YPVESLARNLDWINL MAYDFYGNWSPS QTNSHAQLFDPVNH VSGSDGINAWIQAG VPTKKLVGIFYGYA WRLVNANIHLRAP AAGKSNVGAVDG SMTYNRIRDYIVESR ATTVYNATIVGDYCY SGSNWISYDDTQTV RNKVNYYKGRLLG YFAWHVAGDQNW GLSRTASQTWGVSF QEMK	MANSVTLFAIIFSCFL LQQLVCTNSQNVKG GYWFKDSGLALNNI DSTLFTHLFCFADL NPQLNQLIISPENQD SFRQFTSTVQRKNPS VKTFLSIAGGRANST AYGIMARQPNRSRKS FIDSSIRLARQLGFH GLDLWEYPLSAAD MTNLGTLNEWRTA INTEARNSGRAALL TAAVSNSPRVNGLN YPVESLARNLDWINL MAYDFYGNWSPS QTNSHAQLFDPVNH VSGSDGINAWIQAG VPTKKLVGIFYGYA WRLVNANIHLRAP AAGKSNVGAVDG SMTYNRIRDYIVESR ATTVYNATIVGDYCY SGSNWISYDDTQTV RNKVNYYKGRLLG YFAWHVAGDQNW GLSRTASQTWGVSF QEMK	MANSVTLFAIIFSCFL LQQLVCTNSQNVKGGY WFKDSGLALNNIDSTL FTHLFCFADLNPQLN QLIISPENQDSFRQFTS TVQRKNPSVKTFLSIAG GRANSTAYGIMARQP NSRKSFIDSSIRLARQLG FHGLDLWEYPLSAAD MTNLGTLNEWRTA TEARNSGRAALLTAA VSNSPRVNGLNYPVES LARNLDWINLMAYDFY GPNWSPSQTNSHAQL FDPVNHVSGSDGINA WIQAGVPTKKLVGIFY GYAWRLVNANIHLR APAAGKSNVGAVDG SMTYNRIRDYIVESRAT TVYNATIVGDYCYSGS NWISYDDTQTVRNKV NVYKGRLLGYFAWH VAGDQNWGLSRTASQ TWGVSFQEMK	----- ----- ----LNQL- ISPENQDSFRQ--ST- QRKNP-----RANS TAY--MARQPNSR----- -----P----- -----ARNSGRA- -----PRVNG----- ----- -----PNWSPSQTN--- ----- ----- -----GSMTYNR---Y----- ----- QT----- -----SRTASQ-----	63-66, 68-78, 81-82, 84-88, 100-106, 109- 116, 142, 166-172, 184-188, 213-221, 288-294, 298, 330- 331, 361-366
PR-12	Dahlia	Defensin-like protein 1	ELCEKASKTWSGNC GNTGHCDNQCKSW EGAAHGACHVRNG KHMCFYFNC	ELCEKASKTWSGNC GNTGHCDNQCKSW EGAAHGACHVRNG KHMCFYFNC	ELCEKASKTWSGNCGN TGHCDNQCKSWEGAA HGACHVRNGKHMCF YFNC	-----SKTWS----- DN----- HVRNGKH-----	7-11, 21-22, 36-42
PR-13	Arabidopsis	Thionin 2.1	MKGRILILSLIMSLV MAQVQVEAKICCP NQARNGYSVCRI SKGRCMQVSGCQ SDTCPRGWVNAILE NSADATNEHCKLGC ETSVCGAMNTLQNS DASEIVNGASEQCA KGCSIFCTKSYVPP GPPKLL	MKGRILILSLIMSLV MAQVQVEAKICCP NQARNGYSVCRI SKGRCMQVSGCQ SDTCPRGWVNAILE NSADATNEHCKLGC ETSVCGAMNTLQNS DASEIVNGASEQCA KGCSIFCTKSYVPP GPPKLL	MKGRILILSLIMSLVM AQVQVEAKICCPNQ ARNGYSVCRI MQVSGCQNSDTCPRG WVNAILENSADATNE HCKLGCETSVCGAMN TLQNSDASEIVNGASE QCAKGCSIFCTKSYVPP GPPKLL	----- PSNQARNGY----- -----QNSDTCPR----- ----- NTLQNSDASE----- -----VPPGP-----	29-37, 56-63, 94-103, 126-130
PR-15	Barley	Oxalate oxidase	MGYSKNLGLFT MLLAPAIMATDPD PLQDFCVADLDGKA VSVNGHTCKPMSEA GDDFLFSSKLTAGN TSTPNGSAVTELDVA EWPGTNTLGVSMN RVDFAPGGTNPPHI HPRATEIGMVMKG ELLVGILGSFDSGNK LYSRVVRAGETFVIP RGLMHFQFNVGKTE AYMVVSFNSQNP VFPVPLTLFGSNPPIPT PVLTKALRVEAGVVE LLKSKFAGGS	MGYSKNLGLFT MLLAPAIMATDPD PLQDFCVADLDGKA VSVNGHTCKPMSEA GDDFLFSSKLTAGN TSTPNGSAVTELDVA EWPGTNTLGVSMN RVDFAPGGTNPPHI HPRATEIGMVMKG ELLVGILGSFDSGNK LYSRVVRAGETFVIP RGLMHFQFNVGKTE AYMVVSFNSQNP VFPVPLTLFGSNPPIPT PVLTKALRVEAGVVE LLKSKFAGGS	MGYSKNLGLFTMLL LAPAIMATDPDPLQDF CVADLDGKAVSVNGH TCKPMSEAGDDFLFSS KLTAGNTSTPNGSAV TELDVAEWPGTNTLGV SMNRVDFAPGGTNPP HIHPRATEIGMVMKGE LLVGILGSFDSGNKLYS RVVRAGETFVIPRGLM HFQFNVGKTEAYMVV SFNSQNP GIVFVPLTLFGSNPPIPT PVLTKALRVEAGVVE LLKSKFAGGS	----- DPL----- ----- SSKLTAGNTSTPNG--- -----EWPGT----- ----- -----SGNKL----- -----P----- -----I----- -----	27-29, 62-76, 86-90, 137-141, 181, 196
PR-16	Barley	Oxalate oxidase like or germin like protein	MASSCSFLLAALLA LSWQATSSDPSP QDFCVADMHSPVR VNGFVCKNPMVDN ADDFKAAALDKPR	MASSCSFLLAALLA LSWQATSSDPSP QDFCVADMHSPVR VNGFVCKNPMVDN ADDFKAAALDKPR	MASSCSFLLAALLALV SWQATSSDPSPQDFC VADMHSPVRVNGFVC KNPMDVNADDFKAA ALDKPRVTNKVGSNVT	----- SSDPSPQLQ----- ---KNPMDVN----- AALDKPRVTNKVGS--- -----	23-30, 49-55, 63-76, 135-143, 210-218, 223-226

			VTNKVGSNVTLINV MQIAGLNTLGISIARI DYAPLGQNPPTH RATEILTVLEGTLYVG FVTSNLPAPNRNKF SKVLNKGDFVFPV GLIHQFQNPNPQH AVAIALSSQNP TIANAVFGSDPAISD DVLAKAFQVEKNTID WLQAQFWENNHN	VTNKVGSNVTLINV MQIAGLNTLGISIARI DYAPLGQNPPTH RATEILTVLEGTLYVG FVTSNLPAPNRNKF SKVLNKGDFVFPV GLIHQFQNPNPQH AVAIALSSQNP TIANAVFGSDPAISD DVLAKAFQVEKNTID WLQAQFWENNHN	LINVMQIAGLNTLGISI ARIDYAPLGQNPPTH PRATEILTVLEGTLYVG VTSNLPAPNRNKF VLNKGDFVFPVGLIHF QFNPNPHQPAVAIAL SSQNPGAITIANAVFGS DPAISDDVLAKAFQVE KNTIDWLQAQFWENN HN	----- LPAPNRNKF----- ----- ----- QVEKNTIDW---- FWEN---	
PR-17	Tobacco	NtPrp27	MKKMTEAIFVFSQV KRAHHKIFFFYSFLV AIFTQKIHAVDYSVT NTAANTAGGARFN RDIGAQYSQQTAA ATSIWNTFQQNFP ADRKNVQKVSFV DDMDGVAYASNNE IHVSASYIQGYSQDV RREITGVLYHESHTV WQWNGNGGAPGG LIEGIADYVRLKAGF APSHWVKPGQDR WDQGYDVTARFLD YCNSLRNGFVAQLN KKMRTGYSNQFFID LLGKTVDQLWSDYK AKFRA	MKKMTEAIFVFSQV KRAHHKIFFFYSFLV AIFTQKIHAVDYSVT NTAANTAGGARFN RDIGAQYSQQTAA ATSIWNTFQQNFP ADRKNVQKVSFV DDMDGVAYASNNE IHVSASYIQGYSQDV RREITGVLYHESHTV WQWNGNGGAPGG LIEGIADYVRLKAGF APSHWVKPGQDR WDQGYDVTARFLD YCNSLRNGFVAQLN KKMRTGYSNQFFID LLGKTVDQLWSDYK AKFRA	MKKMTEAIFVFSQV AHHKIFFFYSFLVLAIFT QKIHAVDYSVTNTAAN TAGGARFNRDIGAQYS QQTAAATSIWNTFQ QNFPADRKNVQKVSFV FVDDMDGVAYASNNE IHVSASYIQGYSQDV EITGVLYHESHTVWQ WNGNGGAPGGGLIEGI ADYVRLKAGFAPSHW VKPGQDRWDQGYD VTARFLDYCNSLRNGF VAQLNKKMRTGYSNQ FFIDLLGKTVDQLWSD YKAKFRA	----- ----- GARFNRD-GAQYSQ- ----- QQNFPADRKN----- ----- -----YSGDV----- ----- ----- KPGQGDWRWDQ----- -----M----- RTGYS----- -----	54-60, 62-68, 82-91, 123-127, 176-185, 212-217
PR-18	Sunflower	Carbohydrate Oxidase	MANITSSFNMQTSIL TLTLLSTQSSATSR SITDRFIQCLHADR PSFPITGEVYTPGNS SFPTVLQNYIRNLR NETTTPKPLIATAEH VSHIQAAVVCCKQ RLLKTRSGGHDYEG LSYLTNTNQPFIVD MFNLRNINVDIEQET AWVQAGATLGEVY YRIAEKSNKHGFPAG VCPTVGVGGHFSGG GYGNLMRKYGLSVD NIVDAQIIDVNGKLL DRKSMGEDLFWAIT GGGVSGFVVLAYK IKLVRVPEVTVFTIE RREEQNLSIAERW VQVADKLDRLFLR MTFSVINDTNGGKT VRAIFPTLYLGNRN LVTLNKFPELGLQ ESDCTEMSWVESVL YYTGFPSTPTTALL SRTPQRLNPFKIKSD YVQNPISKRFQEFIF RMKELENQMLAFN PYGGRMSEISEFAK FPHRSGNIAKIQEV NWEDLSDEAENRYL NFTRLMYDYMTFV SKNPRAFLNYRDL IGINSHGRNAYTEG MVYGHKYFKETNYK RLVSVKTKVDPDNFF RNEQSIPTLSS	MANITSSFNMQTSIL TLTLLSTQSSATSR SITDRFIQCLHADR PSFPITGEVYTPGNS SFPTVLQNYIRNLR NETTTPKPLIATAEH VSHIQAAVVCCKQ RLLKTRSGGHDYEG LSYLTNTNQPFIVD MFNLRNINVDIEQET AWVQAGATLGEVY YRIAEKSNKHGFPAG VCPTVGVGGHFSGG GYGNLMRKYGLSVD NIVDAQIIDVNGKLL DRKSMGEDLFWAIT GGGVSGFVVLAYK IKLVRVPEVTVFTIE RREEQNLSIAERW VQVADKLDRLFLR MTFSVINDTNGGKT VRAIFPTLYLGNRN LVTLNKFPELGLQ ESDCTEMSWVESVL YYTGFPSTPTTALL SRTPQRLNPFKIKSD YVQNPISKRFQEFIF RMKELENQMLAFN PYGGRMSEISEFAK FPHRSGNIAKIQEV NWEDLSDEAENRYL NFTRLMYDYMTFV SKNPRAFLNYRDL IGINSHGRNAYTEG MVYGHKYFKETNYK RLVSVKTKVDPDNFF RNEQSIPTLSS	MANITSSFNMQTSIL LLTLLSTQSSATSR RFIQLHADRADPSFPI GEVYTPGNSFPTVLQ NYIRNLRNETTTPKPF LIATAEHVSIIQAAVVC GKQNRLLKTRSGGHD YEGSYLTNTNQPFIV DMFNLRNINVDIEQET AWVQAGATLGEVYRI AEKSNKHGFPAGVCP VGVGGHFSGGGYGNL MRKYGLSDNIVDAQII DVNGKLLDRKSMGED LFWAITGGGVSGFV LAYKILVRVPEVTVF TIERREEQNLSIAERW VQVADKLDRLFLRMT FSVINDTNGGKTVAIF PTLYLGNRNLTLLNK DFPELGLQESDCTEMS WVESVLYTGFPSTPT TTALLSRTPQRLNPFK SDYVQNPISKRFQEFIF ERMKELENQMLAFNP YGGRMSEISEFAKFP HRSGNIAKIQEVNWE DLSDEAENRYLNFTRL MYDYMTFVSKNPRA AFLNYRDLGINSHGR NAYTEGMVYGHKYFK ETNYKRLVSVKTKVDP DNFFRNEQSIPTLSS	----- ATRSITDRF----- DRADPSFP----- YTPGNSS----- LQNYIRNLRNETT----- ----- ----- NVDIE----- ----- LLDRKSM G----- -----EQN----- ---DK----- NDTNGGK----- GNSRN----LNKDFPEL-- QESD CTE----- LSRTPQRLNPFKIKSDY VQNPISKR-----KEL- ----- AKPFPH----- EDLSDEAENRY LNF-----FVSKNPR-- ---RDLI-- NSHGRNAYT--- YGHKYFKETNYK----- ---PDNFFRNEQSI-----	28-37, 43-50, 56-62, 67-81, 144-148, 221- 228, 270-272, 285- 286, 301-307, 319- 323, 328-335, 338- 344, 367-391, 401- 403, 424-429, 444- 457, 468-474, 481- 485, 488-496, 501- 513, 523-533
PR-19	Scots Pine	Anti-Microbial peptide	METKRLAYVMFVLV CLFLAMAQPSQGSY FTAWAGPGCNNHA ARYSKCGCSNIGNN	METKRLAYVMFVLV CLFLAMAQPSQGSY FTAWAGPGCNNHA ARYSKCGCSNIGNN	METKRLAYVMFVLV CLFLAMAQPSQGSY FTAWAGPGCNNHA ARYSKCGCSNIGNN	----- SQGS----- NNHAARYSK----- -----Y-	24-27, 38-46, 72, 74- 88

			VHGGYEFMYQGQT AAAYNTDNCKGVA QTRFSSSVNQACSSF GWKSFFIQC	VHGGYEFMYQGQT AAAYNTDNCKGVA QTRFSSSVNQACSSF GWKSFFIQC	MYQQQTAAAYNTDN CKGVAQTRFSSSVNQA CSSFGWKSFFIQC	TDNCKGVAQTRFSS-- -----	
--	--	--	--	--	---	---------------------------	--

Table 3: Protein Modelling statistics for Swiss Model Portal

PR-Protein	Template used for Homology Modelling	GMQE Score	QMEAN Score	Sequence Identity (in %)	Sequence Similarity (in %)	Query Sequence Coverage (in %)
PR-3	2DKV	0.77	-1.28	66.22	52.00	94.00
PR-9	1QO4	0.79	-1.09	66.11	50.00	93.00
PR-10	5AMW	0.94	0.19	77.85	54.00	99.00
PR-11	3ALF	0.98	-0.97	100.00	61.00	93.00
PR-12	1BK8	0.79	-1.26	69.39	57.00	98.00
PR-15	2ET7	0.96	0.79	98.01	60.00	90.00

Table 4: Protein Modelling statistics for RaptorX

PR-Protein	Templates used for Modelling	P-value	uGDT (GDT) Score	uSeqID (SeqID) Score	Alignment Score
PR-1	5V50, 1SMB, 5JYS, 5ETE, 3Q2R	2.71×10^{-3}	38 (94)	8 (20)	36
PR-4 (Domain 1)	4JP6	2.07×10^{-7}	120 (88)	93 (68)	88
PR-4 (Domain 2)	1HEV, 4MPI, 1K7T	1.52×10^{-5}	42 (62)	43 (63)	35
PR-5	1Z3Q, 1DU5	1.95×10^{-13}	181 (81)	115 (51)	195
PR-6(Domain 1)	1FYB	1.14×10^{-8}	91 (53)	66 (39)	91
PR-6(Domain 2)	1FYB	1.58×10^{-4}	38 (74)	23 (44)	29
PR-7	3VTA	1.43×10^{-10}	269 (74)	140 (38)	240
PR-8	2GSJ, 1KR0, 4TOQ	6.95×10^{-8}	218 (83)	107 (41)	216
PR-13* (Domain 1)	3NIR	2.52×10^{-3}	24 (36)	17 (25)	22
PR-13* (Domain 2)	5FIG, 3LMF, 5ARM, 6EK9, 4WSV	2.08×10^{-2}	17 (25)	10 (15)	9
PR-14 (Domain 1)	2ALG	5.17×10^{-8}	89 (97)	92 (100)	93
PR-14 (Domain 2)	1AFH	1.28×10^{-7}	86 (94)	55 (60)	88
PR-16	2ET7	1.35×10^{-8}	162 (71)	91 (40)	140
PR-17	4AR1, 2Y3U, 4ARE	3.92×10^{-8}	64 (26)	23 (10)	89
PR-18	4UD8, 5D79, 3VTE, 4DNS	5.75×10^{-18}	419 (78)	247 (46)	486
PR-19	1C01	2.56×10^{-5}	69 (66)	51 (49)	69

Table 5a: Docking Interaction Energies of known allergens with the IgE molecule

Family	Source	Allergen	IgE chains interacting with the allergen	Interaction Energy between IgE & the allergen (in kJ/mol)
PR-1	Muskmelon	Cuc m 3	D, E, F, J, K	-369.6
PR-2	Banana	Mus a 5	G, H, I, L	-398.6
PR-3	Chestnut	Cas s 5	H, L	-429.1
PR-4	Rubber	Hev b 6	G, H, I, L	-425.4
PR-5	Banana	Mus a 4	E, F, J, K	-366.1
PR-8	Coffee	Cof a 1	A, B, E, G, I, L	-426.6
PR-10	Apple	Mal d 1	E, F, J	-385.8
PR-14	Peach	Pru p 3	H, L	-486.2

Table 5b: Docking Interaction Energies of unknown allergenicity potential PR-Proteins with the IgE molecule

Family	Source	PR- Protein	IgE chains interacting with the allergen	Interaction Energy between IgE & the allergen (in kJ/mol)
PR-6	Tomato	Proteinase inhibitor	A, G, I, J, K	-545.0
PR-7	Tomato	Pathogenesis related protein P69G	A, B, G, H, I, L	-441.7
PR-9	Tobacco	Lignin-forming peroxidase precursor (EC 1.11.1.7)	H, I, L	-415.4
PR-11	Tobacco	Chitinase, class V	H, I, L	-383.8
PR-12	Dahlia	Defensin-like protein 1	A, E, F, J	-398.2
PR-13	<i>Arabidopsis</i>	Thionin 2.1	A, D, E, F, J, K	-445.8
PR-15	Barley	Oxalate Oxidase	D, E, F, J, K	-383.4
PR-16	Barley	oxalate oxidase-like or germin-like protein	C, D, E, F, J, K	-511.5
PR-17	Tobacco	NtPRp27	A, D, E, F, J, K	-468.1
PR-18	Sunflower	Carbohydrate Oxalate	H, I, L	-478.7
PR-19	Scots Pine	Anti-Microbial Peptide	B, G, H, I, L	-391.5

* The docking energies has been calculated using the equation

$$E = 0.50E_{\text{rep}} + (-)0.20E_{\text{att}} + 600E_{\text{elec}} + 0.25E_{\text{DARS}}$$

Where E_{rep} and E_{att} are Van der Wall's forces of repulsion and attraction, E_{elec} is the electrostatic potential force and E_{DARS} is the statistical potential.

Figure 1

a

b

Figure 2ab

C

PR-15

PR-16

PR-17

d

PR-18

PR-19

Figure 2cd

Figure 3

logo

PR 1 Cuc m 3	RQYANDR.....	9
PR 7 PRp69G	NVYENANGLRPLYDARKAATG.....	56
PR 10 Mal d 1	QISTEGSYGYVK.....	28
PR 12 Defensin like protein	STWSDNH.....	14
PR 3 Cas a 5	..SSPTTTTSSPTQKYRNDPRCKSNQFY..	80
PR 5 Jun a 3	GKRLLDGGQDQYDVSNV.....	53
PR 14 Pru av 3	NNLAITPDNLAI.....	17
PR 18 Carbohydrate oxidase	ATSRISITDRFDRADPSFPYTPGNSSLQNYIRNLRFETITVDIELDRKSMG..EQNKNDTNCKSRPQRLNPKIKSDYVQNPISKELA	93
PR 13 Thionin	PSNQALGYQNS.....	26
PR 11 Chitinase class V	LNQLISPENQDSFRQSTORNPAASTAYMAR.....	69
PR 17 NtPrp27	GARFNRDGAGYSQQQFFP.....	45
PR 8 Cof a 1	GHDEPSSAEFGREGGKSS.....	35
PR 2 Mus a 5	NNLPSPSCAVQSGVTR.....	48
PR 4 Hev b 6	DEYCPDHNCQSNCSGGEHLYSQ.....	57
PR 15 Oxalate oxidase Oxo	DPLSSSLFAGTS.....	30
PR 6 Proteinase Inhibitor	EHAKTKECDTRIPLLETKRVEGKKGCKYFKDSEWVSEKNLKLAKTKECDTRIPLLETKRVEGRKGCKYFSKDEYSEKDNNELEKDCTECD	95
PR 19 Anti microbial peptide	SQGSNINAAYSKYT.....	29
PR 9 Lignin forming peroxidase	RQRDDTDGTOLEKDAFANGRDSLTANRS.....	67
PR 16 Oxo like protein	SDPSLQKQNPMDVIAA.....	50
consensus	*	

logo

PR 1 Cuc m 3	DKGQ.....	9
PR 7 PRp69G		60
PR 10 Mal d 1		28
PR 12 Defensin like protein		14
PR 3 Cas a 5	DCYNQKP.....	87
PR 5 Jun a 3	ATF.....	56
PR 14 Pru av 3		17
PR 18 Carbohydrate oxidase	KFPFHEDLSDEAENRYLNFFVSKNPRDLIDNSHGRNAYTYGHKYFKETNYKRPDNFRNEQSI	157
PR 13 Thionin		26
PR 11 Chitinase class V	SRTASQ.....	75
PR 17 NtPrp27		45
PR 8 Cof a 1		35
PR 2 Mus a 5		48
PR 4 Hev b 6		57
PR 15 Oxalate oxidase Oxo		30
PR 6 Proteinase Inhibitor	TRPLETEKRVKGLGKKGCKYFSEGENDLQKSNV.....	128
PR 19 Anti microbial peptide		29
PR 9 Lignin forming peroxidase	NGQI.....	71
PR 16 Oxo like protein		50
consensus		

X non conserved
 X similar
 X ≥ 25% conserved

Figure 4

Figure 5